HALLUCINATING FACES FROM THERMAL INFRARED IMAGES

Jun Li, Pengwei Hao
Department of Computer Science
Queen Mary, University of London

Chao Zhang, Mingsong Dou
Center for Information Science
Peking University

ABSTRACT

This paper addresses the face hallucination problem of converting thermal infrared face images into photo-realistic ones. It is a challenging task because the two modalities are of dramatical difference, which makes many developed linear models inapplicable. We propose a learning-based framework synthesizing the normal face from the infrared input. Compared to the previous work, we further exploit the local linearity in not only the image spatial domain but also the image manifolds. We have also developed a measurement of the variance between an input and its prediction, thus we can apply the Markov random field model to the predicted normal face to improve the hallucination result. Experimental results show the advantage of our algorithm over the existing methods. Our algorithm can be readily generalized to solve other multi-modal image conversion problems as well.

Index Terms— Photosynthesis, Infrared imaging, Texture synthesis, Image manifold, Graphical Model

1. INTRODUCTION

People identification system based on facial image are welcomed in practice, because they are easy to use for both the examiner and the person being tested, and also because there exists rich reference data for those systems. However, one drawback is that they are easily disturbed by the variance in the sensory data which they are fed, particularly, by uncontrolled illumination. One possible solution is to use thermal sensors to catch the facial image. Long- or middle-wavelength infrared (LW-/MWIR) cameras catch the thermal emission from the subject, which is invariant to external illuminating conditions. Unfortunately existing identity databases mainly consist of normal images, the advantage of rich reference data and human friendliness no longer holds for those IR systems. For example, in the law enforcement, being presented an IR image, most people will feel difficult to recognize a suspect.

To bridge this gap, we develop a learning-based framework, which can "hallucinate" (infer) one's normal (visible-spectrum, VS) looking provided his/her ghost-like thermal facial image. We train the system with pairs of facial images of both modalities (IR and VS). Given an IR face, the system learns what the corresponding VS image should be, by training a canonical correlation analysis (CCA) model. There are difficulties in the way of applying the classical linear CCA on our problem directly: the high dimensionality of the image space and the non-linearity in the possible relations. Therefore, we exploit the local linearity in both the image spatial domain and the image manifolds: For the former, we adopt a patch-based scheme, and for the latter, instead of using the full set of the data to train one CCA, we learn the mapping from a local neighbourhood on the image manifolds. For the global smoothness, the Markov random field (MRF) has been employed to organize the resultant patches as in [1]. In contrast to the previous usage of MRF in image processing, the *observation energy function* in our MRF does not exhibit a trivial definition. We have also proposed a novel measurement of the dissimilarity between the hallucinated (VS) patches and the input (IR) patches to well-pose the objective function of the MRF.

Following a brief review of related work, we analyze the problem and present our framework in Section 2. We conduct comparative experiments in Section 3. Finally, we conclude the paper in Section 4.

1.1. Related Work

Thermal IR facial images have been used for recognition and detection tasks [2]. The combination with VS images have also been explored for robustness and accuracy ([3; 2]). In the previous work, the images of the two modalities are made collaborate in one system, however, the relationship between them remains much unexplored. Relatively few attempts have been made on directly converting between IR and VS faces. Reiter et al. ([4]) have proposed an algorithm, which applies CCA to map images between near IR and VS. However, near IR images capture reflected photons from the subjects in a similar way that the VS images work. Thus these two kinds of images look alike and have shared components, and the conversion is less challenging than that between the VS and the thermal IR images. Linear models have also been used to convert other pair of alike modalities [5].

The relation between a thermal IR image and the corresponding VS image is generally nonlinear. In our early work [6], we exploited the locality in the image spatial domain, in contrast to the holistic models in the previous work. The images are aligned, registered, and cut into small pieces. The linear regression is done on each piece, and pixels in different pieces are taken as independent, which effectively remove much the nonlinear relation we need to consider and make the linear CCA more applicable. In this paper, (i) we further localize our model on the *image manifolds*. (ii) we also take into consideration the relationship between adjacent patches with an MRF. The manifold view of images have been studied ([7]) and applied ([8; 9]) in previous research. Our work can be seen as a new application. We follow the application of graphical model in low-level vision proposed by Freeman et al. [10; 1].

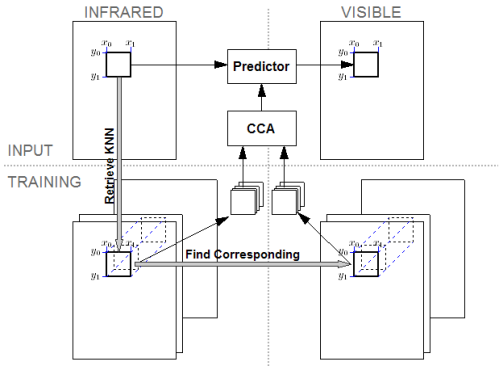

Fig. 1. Flow Chart of Patch Prediction

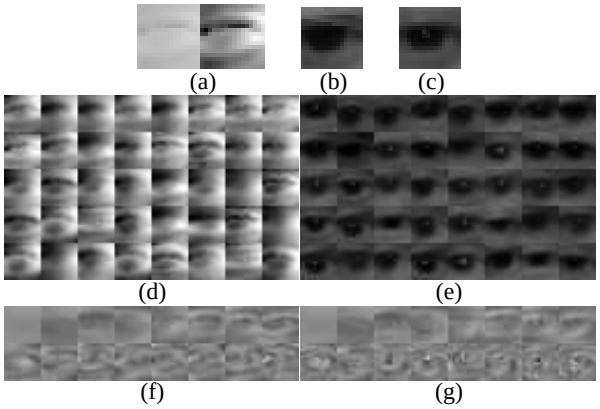

Fig. 2. Predicting VS Patch

(a) Input IR; (b) VS ground truth; (c) Predicted VS; (d) IR nearest neighbors; (e) VS nearest neighbors; (f) IR axes ($\mathbf{Y}_{CC}$); (g) VS axes ($\mathbf{X}_{CC}$)

As we have mentioned, our adoption of the MRF is not trivial, we propose a method to measure the variance between the observations and the predictions.

2. HALLUCINATION MODEL

The thermal emissions of a subject determine its IR image. While the reflective properties underlie its normal photo. Given a certain subject, these two aspects do relate to each other, however, in an obscure way. We try to discover the non-trivial connections by learning from the training IR-VS image pairs. The pairs of facial images have been registered and normalized, and the illuminating conditions for the VS images have been well controlled. This is to eliminate unnecessary variables in the model.

2.1. Locally Linear Correlation

Given the IR patches, linear models are used to predict their corresponding VS patches. We represent an input IR patch as $\mathbf{y} \in \mathbb{R}^D$, where D is its pixel number. Then from each training IR-VS image pair, we cut the patches at the same position and of the same size as $\mathbf{y}$. These IR and VS patches are denoted as $\mathbf{Y}$ and $\mathbf{X}$ respectively. We are to predict $\mathbf{y}$'s VS counterpart $\mathbf{x}$ by using the relations learned from $\mathbf{X}$ and $\mathbf{Y}$.

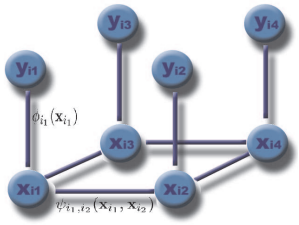

Fig. 3. Markov random field

Following the assumption of image manifolds ([8; 9; 7]), we take $\mathbf{Y}$ and $\mathbf{X}$ as samples drawn from two manifolds. Therefore the neighbourhood of $\mathbf{y}$ in $\mathbf{Y}$ can be seen as lying on a linear subspace. We find the K nearest neighbours in $\mathbf{Y}$ $\mathbf{Y}_{\mathrm{N}}$ for $\mathbf{y}$. Their corresponding VS patches are denoted as $\mathbf{X}_{\mathrm{N}}$. The linear subspaces spanned by $\mathbf{Y}_{\mathrm{N}}$ and $\mathbf{X}_{\mathrm{N}}$ are denoted as $\mathbf{TY}$ and $\mathbf{TX}$, respectively. By introducing the use of neighbourhoods, $\mathbf{X}_{\mathrm{N}}$ and $\mathbf{Y}_{\mathrm{N}}$, we improve the applicability of the linear model from the previous work [6].

We use CCA to model the linear relations between $\mathbf{TY}$ and $\mathbf{TX}$. CCA finds one set of axes for each dataset, along which these two sets of data co-vary most [11]. In the viewpoint of learning, CCA finds the most linear predictable components for the two sets. Formally, if $\mathbf{y}_{CC}^1$ is the first *canonical correlation (CC) axis* of $\mathbf{TY}$, and $\mathbf{x}_{CC}^1$ is that of $\mathbf{TX}$, then they are found by maximizing:

$$\mathbf{y}_{CC}^1, \mathbf{x}_{CC}^1 := \underset{\substack{\mathbf{u},\mathbf{v} \in \mathbb{R}^D \\ \|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1}}{\operatorname{argmax}} \mathbf{u}^T \mathbf{Y}_{\mathrm{N}} \mathbf{X}_{\mathrm{N}}^T \mathbf{v} \qquad (1)$$

The k-th axes $\mathbf{y}_{CC}^k$ and $\mathbf{x}_{CC}^k$ are found in the same way, with the constraint of being perpendicular to the subspace spanned by the first $(k-1)$ axes. Solutions to these axes is can be converted to an eigen-problem[12]. Having computed the CC axes $\mathbf{Y}_{CC} = \{\mathbf{y}_{CC}^k, k = 1, \ldots, D_{CC}\}$, we obtain a subspace of $\mathbf{TY}$, $\mathbf{TY}^{CC} = \mathrm{span}\{\mathbf{Y}_{CC}\}$, such that by projecting $\mathbf{Y}_{\mathrm{N}}$ into $\mathbf{TY}^{CC}$, the obtained feature vectors preserves most of the information about $\mathbf{X}_{\mathrm{N}}$ [13].

Then the regression matrix $\mathbf{M}$ is:

$$\mathbf{M} = (\mathbf{Y}_{CC}^T \mathbf{Y}_{\mathrm{N}})^+ \mathbf{X}_{\mathrm{N}} \qquad (2)$$

where $+$ represents pseudo-inverse. The prediction is:

$$\mathbf{x} = \mathbf{M}\mathbf{Y}_{CC}^T \mathbf{y} \qquad (3)$$

In Figure 1, we draw a flowchart of the procedure of the prediction. In Figure 2, we show a concrete example: For the IR input patch $\mathbf{y}$ (a), 40 nearest neighbours (d) and their corresponding VS patches are found (e). They are $\mathbf{Y}_{\mathrm{N}}$ and $\mathbf{X}_{\mathrm{N}}$ respectively. 16 CC axes are computed for both $\mathbf{Y}_{\mathrm{N}}$ (f) and $\mathbf{X}_{\mathrm{N}}$ (g). Then the IR samples in (d) is projected into the space $\mathbf{TY}^{CC}$ (spanned by the canonical axes showed in (f)). The regression matrix $\mathbf{M}$ is then computed as Eq(2). Then the prediction in (b) is computed as $\mathbf{M}\mathbf{Y}_{CC}^T\mathbf{y}$.

Note that $\mathbf{X}_{\mathrm{N}}$ and $\mathbf{Y}_{\mathrm{N}}$ are assumed to be centered at origin.

2.2. Markov Random Field of Patches

After obtaining the predicted VS patches, we further improve the result by adjusting the VS patches for the global smooth-

ness of the image. As in [10] we use an MRF to enforce the "agreement" between adjacent patches on the overlapped area.

An MRF is a graphical model as shown in Figure 3, where IR and VS patches are $\mathbf{y}_\bullet$ and $\mathbf{x}_\bullet$ respectively. Connections are made between each pair of IR-VS patches and adjacent predicted VS patches. The joint probability of a given set of input IR patches and the predicted VS patches is determined by the energy on those connections:

Observation energy: It measures the deviation of the predicted $\mathbf{x}_i$ from the observation $\mathbf{y}_i$. Different from the previous application of MRF on low-level vision tasks, in our problem, there is no intuitive way of measuring this variance. Given a VS patch $\mathbf{x}_i$ at node i, we propose to do the inverse mapping and measure the variance in $\mathbf{T}\mathbf{Y}_i^{CC}$ by:

$$\log \phi_i(\mathbf{x}_i) := \frac{\|\mathbf{M}_i^{+}\mathbf{x}_i - \mathbf{Y}_{iCC}^{T}\mathbf{y}_i\|_2}{\sigma^{\phi}} \tag{4}$$

where $\mathbf{Y}_{iCC}^{T}\mathbf{y}_i$ is the projection of $\mathbf{y}_i$ in $\mathbf{T}\mathbf{Y}_i^{CC}$. In optimization, it favors adjustments that keep the VS patch related to the input: Consider two adjustment vectors $\Delta\mathbf{x}_1$ and $\Delta\mathbf{x}_2$ with equal norms. Then the cost of these two adjustments in terms of the observation energy in Eq(4) shows that the adjustment that changes less correlated components costs less observation energy.

Transition energy: It puts the smoothness constraint. If ($\mathbf{x}_i$ and $\mathbf{x}_j$) share pixels, and the pixels in the common area are indexed by α_{ij} in $\mathbf{x}_i$ and α_{ji} in $\mathbf{x}_j$, the transition energy is:

$$\log \psi_{i,j}(\mathbf{x}_i, \mathbf{x}_j) := \frac{\|\mathbf{x}_i(\alpha_{ij}) - \mathbf{x}_j(\alpha_{ji})\|_2}{\sigma^{\psi}} \tag{5}$$

In the optimization, when one patch is updated, all other patches are kept fixed. It is local and fast. However, we can further lower the dimension of the space in which we search the optimal solution. We apply principal component analysis [14] on $\mathbf{X}_{i\mathrm{N}}$, and limit our search within the subspace spanned by D_{opt} top eigenvectors. Thus in each step, it is to to find the optimal D_{opt}-dimensional vector $\mathbf{w}_i$:

$$\mathbf{x}_i \leftarrow \mathbf{x}_i^0 + \mathbf{X}_{iPC}\mathbf{w}_i \tag{6}$$

in order to minimize the energy functions

$$\mathrm{Log}E_i = \log \phi_i(\mathbf{x}_i) + \sum_{j,\mathbf{x}_i \sim \mathbf{x}_j} \log \psi_{i,j}(\mathbf{x}_i, \mathbf{x}_j) \tag{7}$$

where $\mathbf{x}_i^0$ is the initial CCA prediction, $\mathbf{X}_{iPC}$ is the first D_{opt} principle components. The optimization in subspace is not only fast, it also limit $\mathbf{x}_i$ within the primary varying directions of $\mathbf{X}_{i\mathrm{N}}$ and thus is more robust.

Note that only the ratio of the normalization parameters σ^{ϕ} and σ^{ψ} actually counts, they are chosen according to other experiment parameters. Our algorithm is summarized in Algorithm **1** and **2**.

Algorithm 1 Local Prediction

1: **for** each i-th patch **do**
2: $\mathrm{N} \leftarrow K$ nearest samples $\in \mathbf{Y}_i$ of the input IR patch $\mathbf{y}_i$
3: Compute $\mathbf{Y}_{iCC}$ and $\mathbf{X}_{iCC}$ from $\mathbf{Y}_{i\mathrm{N}}$ and $\mathbf{X}_{i\mathrm{N}}$
4: Compute $\mathbf{M}$ in Eq(2)
5: Initial prediction $\mathbf{x}_i^0 \leftarrow \mathbf{M}\mathbf{Y}_{iCC}^{T}\mathbf{y}_i$
6: Compute first D_{opt} principle components $\mathbf{X}_{iPC}$ for $\mathbf{X}_{i\mathrm{N}}$
7: **end for**

Algorithm 2 Optimization

1: Generate a random visiting queue $\mathbf{Q}$ of the patches
2: **while** $\mathbf{Q} \neq \emptyset$ **do**
3: Remove the first patch i-th patch from $\mathbf{Q}$.
4: Optimize $\mathbf{w}_i$ minimizing $LogEnergy_i$ as in Eq(6) and Eq(7)
5: **if** $\mathbf{w}_i$ changed **then**
6: Add adjacent patches in Q, if they have not been yet.
7: **end if**
8: **end while**

3. EXPERIMENTAL RESULTS

In our experiments, our algorithm is compared with several developed methods of regression and learning-based texture synthesis. We use the publicly available database ([15]). From the database, we take 20 pairs of IR and VS images for each of the 47 subjects to train our algorithm. Images are preprocessed as abovementioned. The leave-one-out scheme is used, for each input IR face, all the IR-VS pairs of other subjects are used as training data.

The parameters are chosen as follows: (i) Patch size: Larger patches are more likely to capture the local features, while smaller ones are easier to predict. Our algorithm works robustly in a reasonable range (Figure 4). Generally, we use 9×9, with 3-pixel overlapping. (ii) Neighbourhood size K: Large K allows rich samples. Small K makes the local tangent space be better linearly approximated; (iii) Number of CC axes D_{CC}: Smaller D_{CC} tends to give more robust prediction. More axes, on the other side, allow finer control. K and D_{CC} affect the CCA prediction collaboratively. The result at different combination is shown in Figure 5. Our algorithm is robust in a wide range (Figure 4). Generally, we use $K = 45$ and $D_{CC} = 10$. (iv) $D_{opt} = 3$ is used for higher optimization speed. (v) σ^{ϕ} and σ^{ψ} are set to 1.

In Figure 6, the results of our method on middle-wavelength IR (MWIR) images are compared to that of holistic CCA/Eigen-Transformation [4][5], patch-based LLE [8] and the MRF model[10]. Our method yields the most visually plausible

(a) (b) (c) (d) (e) (f) (g)

Fig. 4. Parameter Variation Test

(a) Input IR; (b) Ground truth; (c)-(g) Hallucination results with different parameters:

	K	D_{CC}	PatchSize	Overlap
(c)	45	10	7	2
(d)	45	10	15	5
(e)	15	8	9	3
(f)	45	10	9	3
(g)	85	15	9	3

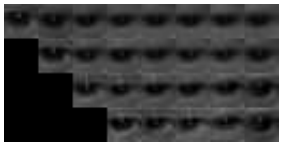

Fig. 5. Effect of K and D_{CC}
K from left to the right: $10, 20, \ldots, 80$; D_{CC} from top to bottom: $8, 16, 24, 32$.

(a) (b) (c) (d) (e) (f)
Fig. 6. Photo-realistic Face Synthesis
(a) Input MWIR; (b) Predicted by holistic CCA; (c) LLE; (d) MRF[1]; (e) Ours; (f) Ground-truth

results. MWIR images are used for comparison because they are of higher quality than that of long-wavelength IR (LWIR) images. For the comparison, we want to eliminate as much as possible influences from irrelevant variables such as the noise. Results of our algorithm on LWIR images are shown in Figure 7.

We conduct a simple recognition test with a "eigenface + K-NN" [14] classifier. In each test, we compare one synthesized VS face with the VS faces, and record if it has been correctly recognized. The recognition result is listed in Table 1. Although the recognition rate is not satisfying, compared to other algorithms, our algorithm's resulting images are more distinguishable.

4. CONCLUSION

In this paper, we proposed a learning based framework to address the new problem of hallucinating facial images from thermal infrared images. For the model's generalization ability, our algorithm works locally both in the sense of image spatial domain and on the image manifold. We use an MRF model to organize the patches to put smoothness and alikeness constraints. We also propose a metric to measure the alikeness between an IR patch and its VS result. At each objective

Table 1. Recognition Rates

Method	$D = 8$	$D = 16$	$D = 24$	$D = 32$
CCA	6.38	4.25	4.25	6.38
LLE	17.02	19.14	25.53	31.91
MRF	14.89	23.40	23.40	23.40
Ours	**19.14**	**40.42**	**44.68**	**50.06**

D is the eigenfaces used.

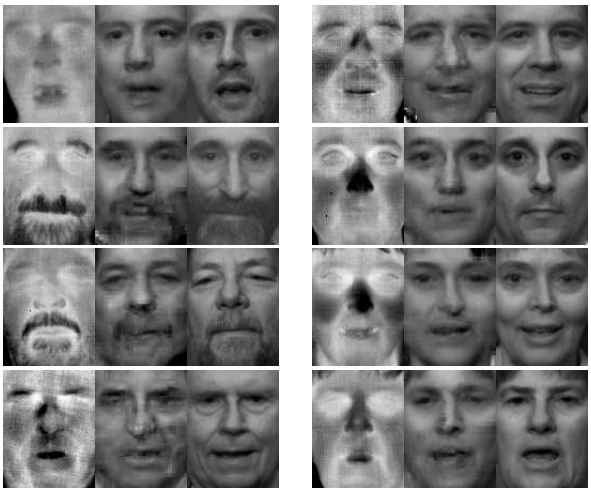

Fig. 7. Hallucination from LWIR images
Left: LWIR; Middle: Hallucinated; Right: Ground-truth.

patch in the MRF, we adjust it in a subspace of its tangent space to the manifold. Such that adjustments in optimization robustly result in meaningful image patches. The framework is effective and robust. It performs satisfactorily in our testing experiments. Comparative experiments demonstrate the better performance of our algorithm as well.

Future work should include generalizing the algorithm to other multi-modal image conversion or synthesis tasks and making the model more robust to ill-registered training image pairs.

References

[1] W.T. Freeman, T.R. Jones, and E.C. Pasztor, "Example-based super-resolution," *IEEE Computer Graphics and Applications*, 2002.

[2] S.G. Kong, J. Heo, B. R. Abidi, J. Paik, and M. A. Abidi, "Recent advances in visual and infrared face recognition – a review," *Computer Vision and Image Understanding*, 2005.

[3] K.W. Bowyer, K.I. Chang, P.J. Flynn, and X. Chen, "Face recognition using 2-d, 3-d, and infrared: Is multimodal better than multisample?," in *Proceedings of the IEEE*, 2006.

[4] M. Reiter, R. Donner, G. Langs, and H. Bischof, "3-D and infrared face reconstruction from RGB data using canonical correlation analysis," in *ICPR*, 2006.

[5] X. Tang and X. Wang, "Face sketch recognition," *IEEE Trans. Circuits Syst. Video Techn*, 2004.

[6] M. Dou, C. Zhang, P. Hao, and J. Li, "Converting thermal infrared face images into normal gray-level images," in *ACCV*, 2007.

[7] D. Beymer and T. Poggio, "Image representation for visual learning," *Science*, 1996.

[8] H. Chang, D.Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *CVPR*, 2004.

[9] Wei Fan and Dit-Yan Yeung, "Image hallucination using neighbor embedding over visual primitive manifolds," in *CVPR*, 2007.

[10] W.T. Freeman and E.C.Pasztor., "Learning low-level vision," *IJCV*, 2000.

[11] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.

[12] T. Kim, J. Kittler, and R. Cipolla, "Learning discriminative canonical correlations for object recognition with image sets," in *ECCV*, 2006.

[13] Magnus Borga, "Canonical correlation a tutorial," 2001, hppt://people.imt.liu.se/ magnus/cca/.

[14] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, 1991.

[15] Diego A. Socolinsky and A. Selinger, "A comparative analysis of face recognition performance with visible and thermal infrared imagery," in *ICPR*, 2001.